Clinical Chemistry

Postmarket Surveillance of Point-of-Care Glucose Meters through Analysis of Electronic Medical Records

Article:
Lee F. Schroeder, et al.
Postmarket Surveillance of Point-of-Care Glucose Meters through Analysis of Electronic Medical Records.
Clin Chem 2016;62:716-724.
http://www.clinchem.org/content/62/5/716.abstract

Guests:
Dr. Lee Schroeder is Assistant Professor and Director of Point-of-Care Testing in the Department of Pathology at the University of Michigan School of Medicine.

Bob Barrett: This is a podcast from *Clinical Chemistry*, sponsored by the Department of Laboratory Medicine at Boston Children's Hospital. I'm Bob Barrett.

Postmarket surveillance is an important monitor of the safety of pharmaceuticals and medical devices after regulatory approval and entry to the market. Although premarket studies are substantial, costs prohibit examination of heterogeneous populations or rare events over long durations. The U.S. Food and Drug Administration monitors postmarket surveillance of in vitro diagnostics usually through passive adverse event reporting.

However, passive reporting often includes errors or omissions, insufficient descriptions of incidents, incorrect product codes and miscategorization of adverse events. In addition, passive reporting usually does not include quantitative data such as trueness and precision. The electronic medical record, or EMR, holds a promising source of data for active postmarket surveillance of diagnostic accuracy, particularly for point-of-care devices.

The May 2016 issue of *Clinical Chemistry* published a paper that compared prospective bedside and laboratory accuracy studies that demonstrated the validity of active surveillance via an EMR data mining method comparing point-of-care glucose results to near in-time central laboratory glucose results.

Dr. Lee Schroeder is the senior author of that paper and he joins us in this podcast. He is Assistant Professor and Director of Point-of-Care Testing in the Department of Pathology at the University of Michigan School of Medicine. Dr. Schroeder, what were the origins of this study and why did you undertake the project?

Dr. Lee Schroeder: To understand, I think, why there is so much interest in the accuracy of a single test, we need to consider history a little bit. In particular, it's use in strict glycemic control of critically ill patients. These patients tend towards

hyperglycemia in the ICUs and hyperglycemia can cause significant morbidity, and about 10 or 15 years ago, there were a series of observational and randomized controlled trials showing that strict glycemic controls and keeping these patients in a normal range of glucose reduce mortality sometimes significantly.

And so, then there was widespread adoption of this protocol in ICUs, what happened was that larger trials started coming out and showing that there wasn't benefit. Now, there were a number of differences in design between the early trials and the late trials, but one of the differences was that they used different instruments to measure the glucose. So in this initial study they were using sophisticated blood gas analyzers and in that widespread adoption, the protocol is typically implemented with the handheld glucose meters. What's important is that these glucose meters were never approved by the FDA for use in those populations, which cast some doubt on those meters' suitability and accuracy.

On top of all this, at that same time period, there was another development which was the maltose interference. Some devices were being reported to the FDA as being associated with mortality in patients who are receiving maltose-based medicine. The specific one was the peritoneal dialysis fluid which contains maltose. So these patients are being checked for glucose and the maltose is making the glucose meter read very high. So they were getting a lot of insulin, and being driven into dangerously low hypoglycemic states.

And so, I think it was because of these two things that the FDA released draft guidance a couple of years ago to test manufacturers, describing the requirements of accuracy they would need to see before they approve the meter for use in the critically ill. And they shortly actually approved one meter after and that's the only meter that's been approved today. And then soon after, CMS released a memorandum reminding everyone who was using other meters, the meters that weren't approved for the critically ill, that if they were to use those meters, it would be considered a laboratory-developed test, off-label use, and therefore high complexity. When a test is high complexity, it requires a whole bunch of hurdles, regulatory hurdles, including a validation of the device in your institution in that in your patient population of interest.

That caused ripples throughout the testing community. Some health systems just decided to drop their glucose meters and go with the one that was recently approved. But health systems can have 300 or 400 or 500 or more of these meters, so that's quite an expensive proposition.

That's sort of the backdrop of the story. And the glucose meter, the accuracy is under serious scrutiny particularly in the critically ill patients.

I think it's also important to talk about what systems are in place, at least in the US, to ensure safety of diagnostic devices. Of course, the first is the FDA. They require a premarket approval. And so, in the case of the meter that was approved for the critically ill, the FDA required a 1,650 patients study to ensure that the meter was accurate in that population.

That's a large study. Most accuracy studies, they're often 200 or 300 patients. But it's also not large enough to account for the great heterogeneity of patients you're likely to see once the meter is release onto the market. In itself, it's not really enough to assess the interferences from a myriad of other drugs that patients are going to be seeing, specifically in the ICU where polypharmacy is the norm, to know is there another maltose on the horizon.

After the premarket approval, there is postmarket surveillance. What the FDA provides is what's called passive reporting or adverse event reporting. There's a database that you can report events where you think the device was associated with some morbidity or mortality. And that's of course is very useful, but it doesn't capture everything and it also doesn't give you a sense of the accuracy of the instrument.

Another system is a mechanism called proficiency testing, which is a regulatory activity. There are a number of proficiency testing providers in the United States. And so, the idea do is that they will send out a sample that contains a certain amount of glucose out to all their participants, and they might have 40,000 participants in their group. And then each lab will run their point-of-care meter on the sample and report the results and then get graded. Theoretically, it could be an excellent postmarket surveillance. You should be able to tell when some meters aren't performing well and others are.

The problem is point-of-care meters require whole blood. But whole blood had red cells that continue to metabolize glucose. So you can't send whole blood out to all these labs because the glucose level would be different by the time that the sample has got to them. So instead, you have to send an artificial sample filled with various chemical constituents that look kind of like blood but not quite. Unfortunately, those interfere with some devices and not other devices. So you don't know if the problem is with the meter or with the sample.

I'm actually on the Chemistry Resource Committee of the College of American Pathologists. And so, we review and recommend, I make recommendations to the college on their proficiency testing program. To deal with this, we wanted to each go back to our home institutions and run our point-of-care glucose meters on a sample, and they run that same sample on our central laboratory instrument, get a sense of the accuracy and then pool the results and get sort of multi-center picture of accuracy of the different meters.

When I went back, at that time I was at Stanford, I was a resident and I went to my point-of-care testing manager and asked her to run a few samples. And she told me, "Well, every once in a while, a couple of times a year I will sift through our point-of-care results coming from the ICUs and then find some events where they happen to also have sent some to the lab. And then I can just look them up if they're near in time and do my little accuracy study that way." And so, once that sunk in, I understood that there could be an enormous number of these so-called "coincident events" in the medical record just waiting to be uncovered if would do it in a comprehensive way.

So we did that in Stanford. And actually, I have since moved to the University of Michigan so I had to redo the entire study. But here we found 27,000 events over a 21-month period in approximately a minus-five-minute time window which is by an order of magnitude larger than these large studies going to the FDA. What's exciting about that is when you go up a scale like that, you should be able to do some much more interesting studies.

Bob Barrett: What informatics or data mining methods did you use and are they available to other researches?

Dr. Lee Schroeder: This method is based on this concept of coincident testing. So if there's a patient who received a point-of-care glucose reading, and near in time to that, they happen to have been drawn, the blood was drawn and then sent to central lab also for glucose testing, that event is an opportunity to learn something about the accuracy of the point-of-care meter. I think this falls under this idea of the Learning Health System, which is much of healthcare has been based on the controlled and randomized trials, but they're expensive. And so, there's a greater focus now on observational trials where you take advantage of the thousands of natural experiments that are occurring each day in a hospital system. But if you don't pool them together, then you can't learn from them. So that's what we're trying to do here.

Now, the coincident testing might happen out of convenience. A nurse may need to run in the ICU their hourly point-of-care glucose. Well, they also have to draw

blood for basic metabolic panel. So it's more convenient for them to do it at the same time. So often if they're going to do it at the same time, they take blood from arterial or venous line and use that same blood in both point-of-care meter after the central lab. And that's good for us because then we're comparing apples to apples. The point-of-care meter is being run on the same sample that the central lab is being run on.

Now, the problem with this approach, of course, is that if all you know is that these two tests were collected in the same period of time or the blood was collected in the same period of time, you don't know if something happened. Maybe there's like five minutes difference in time. You don't know if there was insulin or glucose administration which could cause increase in value in one that gets the second measure in that case. That's the thing that we're worried about, and so we developed a set of filters to try to clean up these coincident events to select the best datasets.

Anyway, to do this, we needed a gold standard. We were going to have the data mining approach which is going to give us some estimate of accuracy or bias and a random error of instrument, but we needed to know what the actual bias and random error was. And we saw the best way to do that was to run a prospective study in the ICU, recruit nurses. And so, every time they in fact drew blood from an arterial line and used that same blood on both the point-of-care meter and then sent some down to the central lab, they would flag that event on the meter and then we would be able to collect those. That was our gold standard.

And we implemented a large number of filters. And in the end only a few of them were really important, which is the good news, I think, for anyone who wants to implement this. But we of course filtered on the time of collection and we had to first -- well, it turns out timestamps in the electronic medical record do not actually represent time. For instance, if you're collecting blood for essential lab measurement, the time of collection is actually the time when the label is printed for the tube, not the time when you're drawing blood from the patient. So we had to make adjustments for that.

We also had to the fact that when you send blood to the central lab, there is continued glycolysis. The red cells continue to metabolize the glucose. And so, there's going to be this positive bias that will appear to be a positive bias in the point-of-care meter if you don't account for that. So we had to do some regressions to figure that out. And then location is important. The ICU, that's where most patients will have lines, and arterial lines would be the best. You don't have a concern for IV contamination of IV fluids. But

since more patients are going to have arterial lines, they're less likely to get a fingerstick for their point-of-care testing reading. And we don't want to compare fingerstick measurements to arterial measurements. So we also filtered for ICU location.

And then the other important one was time period. The bedside ICU study occurred over about five months. And so, we restricted the study, we filtered a six-month period because we didn't want to take into account long-term variation, maybe changes in the reagent last and so forth that we didn't have confounding the ICU study.

And so, in the end, we end up with about 27,000 events when you filter just on the time of collection and with successive filtering, actually, the most sophisticated filtering we did was to get actual line placements of the patients from the medical record. So that was the only piece of information we needed outside the laboratory information system. And when we implemented that, we came down to 852 events. What we found was that the method corresponded very well, with the bedside ICU study. So the bias was in very good agreement. And essentially, it was non-existent after we accounted for the glycolysis. They're non-existent in both the bedside ICU study and in DETECT which is our algorithm. And the random error was also in very good agreement, about 5% or 6% between the two methods.

The other approach people sometimes use is to look for how many or what percent of events are outside of some quality goal. So if you'll assign the quality of these meters as requiring 95% within 12.5% of the gold standard when glucose is over 100 mgs per deciliter. The bedside ICU study is on 2.4%, falling within that goal, and DETECT is about 4.8%. It's a little bit larger, but that's the cost you're going to pay to get these large datasets, which is that you're going to get some outliers. So on this case, we had 8 outliers out of our 852 samples. When I looked into that, four of them could be written off, three of them from IV contamination, and one was a point-of-care test that was repeated two or three minutes later that was normal. So there might be a way to reduce those outliers, but that's just I think something you're going to have to live with when you want to get large datasets like this.

In the end, these findings validated this method as reliable for evaluating accuracy of glucose meters. And one thing I didn't mention was -- this is not my idea -- this practice is somewhat widespread, and I think it's done manually quite often but more and more is it being done in an automated fashion. But because it's never been validated, we thought it was important to publish this paper.

Bob Barrett: And how is this type of study going to help other clinical laboratories?

Dr. Lee Schroeder: I think there will be a couple very practical applications. First of all, if you are a hospital system that's going to use a glucose meter or continue to use your glucose meter in the critically ill and that meter has not been approved for the critically ill, if you do need to run a validation study, you could use this approach and gain a very good confidence of its accuracy within the critically ill.

Another regulatory requirement is doing intra-instrument correlations. So if your lab provides testing for glucose on multiple different types of instruments, you're supposed to twice a year ensure that those different instruments are resulting in a similar way so there's some harmonization across your institution. So at best, people are just looking at a subset of their instruments, well, I guess the analyzers, glucose meters, central lab analyzers. By using this method, you can cast a much wider net and get closer to a total quality assurance.

Now, I think to comment on how hard this is, I think that's a question I'll get. Importantly, we show that most of the improvements in the algorithm, all those filterings that I mentioned, most of the improvement can be achieved with relatively simple filtering. There's a just few steps just with data from the laboratory information system.

I think it's not that complicated of a method but it's not something you can do in like worksheet like Excel. So I was working in our statistical environment, and I think to make this happen you need to do that. Or you could possibly do this and the past informatics group could do this as part of a sequel query and maybe that would be the best way to implement it.

We spent a fair amount of time figuring out what the time delay was or the translation of the timestamps in the medical records to the actual times on the floor, but in the end, I went back and reran it, assuming there was no delay. So just a plus or minus five minutes around the point-of-care measurement, I also expanded it from zero to 10 to go to minus-6 to 16 and all those results looked essentially the same. So, potentially you would not have to do something like a perspective bedside ICU study to verify what your timestamps are going to be and so forth. In other words, this method appears to be quite robust.

Bob Barrett: Are the procedures your group used different from other methods for accuracy assessment?

Dr. Lee Schroeder: I think this is where we can start talking about what you can do with this much larger dataset. In safety studies of pharmaceutical, pharmacovigilance, there has been quite a number of publications, they've done interesting things looking at uncovering novel adverse events or off-label usage patterns or interactions, new interactions between drugs. You know, it's in that spirit that I think we can start looking at diagnostics if we start mining into the medical record. So of course, we can use it for simple accuracy studies for bias and random error. With these larger datasets, we'll get a much more granular understanding. But we can also use it for novel studies that are only enabled with large datasets.

Now, with respect to the strict glycemic control and issues in the critically ill, I believe it's the polypharmacy and the extremes of physiology that are of greatest concern. Now, with this much data, what we should be able to do is regress the error found in all these coincident events against pharmaceutical administrations and against extremes in physiology or other laboratory values and look to see if there are some patterns. And maybe we find there is none. And so, the concern for the next maltose is unfounded.

We did it as a proof of principle looking at lab results in the CBC and the basic metabolic panel. There, we took 17,000 of our events and did a multivariable regression against those lab analytes, and first of all confirmed that the largest interference is with hematocrit which is well known and it was at similar order of magnitude. And also showed that really, while there were some other associations, they weren't clinically relevant.

I think it's important that as we applied more and more filters to clean the data up, we got continued improvement or a closer correlation of the tax estimation of random error with the bedside ICU study. But even with just a couple of filters, the bias was dead on with the bedside ICU study. And it's bias that you expect to be introduced when you get interferences with drugs or other labs. And so, one should be able to use the largest dataset to look for interferences.

Another advantage to the approach is that in a smaller study, you're rarely going to find patients with an extreme glucose value. So a patient that has very severe hypo or hyperglycemia. And because of that, the FDA allows companies to spike their samples with the glucose or to reduce glucose levels. So a common method is to let your sample sit at room temperature and let the red cells metabolize the glucose. And that's the best we can do, but it's somewhat artificial. And so you are always concerned that maybe that's impacting the results somehow.

But when you have 27,000 events, that opens the possibility of finding these hypoglycemic patients in real-time and in vivo. And so, we should be able to get a better understanding of how these instruments work at these extreme values of glucose.

Bob Barrett: Well, finally, doctor, is this methodology restricted to examining point-of-care glucose devices or can it be used for other tests in other settings?

Dr. Lee Schroeder: There is a great pressure in the U.S. health system to improve efficiency of care and that's one of the reasons point-of-care testing is becoming so popular now. The idea is that if you use a point-of-care test you won't have to wait for a laboratory results, a central laboratory result, you would be able to change patient management immediately. Of course, the central laboratory is able to, if their economy is at scale, use very sophisticated instruments with excellent accuracy. In point-of-care testing, that's the biggest concern, that you've got an accuracy issue.

So if we can use this method on other point-of-care tests and allow us to detect when point-of-care instruments are in fact not operating with sufficient accuracy, I think you can go a long way in successful implementation of point-of-care testing in the health system. So glucose is by far the largest point of care test with the largest volume. So in our institution, we get about 5,000 tests weekly. The other point-of-care tests are about 500. So it's much lower. But with the glucose, we're starting out with 27,000 events. So even if you're down an order of magnitude, you're looking at a couple of thousand events. And so, that's plenty, right? And good published accuracy studies are often in the hundreds.

Now, it remains to be shown if the method is valid for other analytes. I haven't done any prospective studies as a gold standard to validate it against other point-of-care tests, but I've got some anecdotal and some empirical evidence that it seems to be working fine. For instance, our troponin testing that we do, we knew we validated the troponin test that there was a 50% negative bias in the point-of-care instruments as compared to our central lab. And that's fine because that's not an analyte, that's standardized or harmonized and you expect there to be differences. So when I did these coincident testing analyses, we found the exact same thing, that there is this 50% negative bias. But we also found because we have so many data points that that normalized at low levels, which is something we didn't realize before.

The A1C was interesting because there was actually more variation there than I was expecting. So that's something I

want to look into more closely. We did this with a point-of-care lipids test and found what we had expected in validation studies, which was that the total cholesterol is pretty good but the HDL showed about a third negative bias at low levels.

And then recently, I turned this around. So we had a request from a physician offsite to look into the potassium. He felt like he was getting more inappropriately high or artificially high potassium results in the central lab. And so he would have the patient come in and get a point-of-care potassium test just to make sure the patient wasn't high. And he thought this was a new development.

So I was able to turn this around and actually use a point-of-care test as the reference and then look to see if there was error, most certainly due to pre-analytic factors like release of potassium from the cells on their trip to central lab and showed that, in fact, there was no systematic change in the results and that they matched fairly well with the point-of-care testing. So probably, it's just an issue with one particular patient.

I think where this is so useful to me for is that we get concerns all the time from providers saying, "Oh, I don't trust this instrument. We got this very abnormal result a couple of months ago and I think you need to look into it." But they won't have a name, a medical record number. And so, it's like this he said/she said thing. Now we can quickly mine the data, find compared events, coincident events, and then at least have something on paper to talk about and set us in the right direction.

Bob Barrett: Dr. Lee Schroeder is Assistant Professor and Director of Point-of-care Testing in the Department of Pathology at the University of Michigan School of Medicine. He has been our guest in this podcast from *Clinical Chemistry* on postmarket surveillance of point-of-care glucose meters to analysis of electronic medical records. His paper on that topic appeared in the May 2016 issue of *Clinical Chemistry*.

I'm Bob Barrett. Thanks for listening!